

Managing large-scale scientific hypotheses as uncertain and probabilistic data with support for predictive analytics

Bernardo Gonçalves^{1, a)} and Fabio Porto^{1, b)}

¹⁾ *Computer Science Dept., National Laboratory for Scientific Computing (LNCC)*^{c)}

(Dated: 19 May 2015)

The sheer scale of high-resolution raw data generated by simulation has motivated non-conventional approaches for data exploration referred as ‘immersive’ and ‘in situ’ query processing of the raw simulation data. Another step towards supporting scientific progress is to enable data-driven hypothesis management and predictive analytics out of simulation results. We present a synthesis method and tool for encoding and managing competing hypotheses as uncertain data in a probabilistic database that can be conditioned in the presence of observations.

Keywords: hypothesis management, predictive analytics, synthesis of probabilistic databases.

^{a)}Electronic mail: bgonc@lncc.br.

^{b)}Electronic mail: fporto@lncc.br.

^{c)}<http://dexl.lncc.br>.

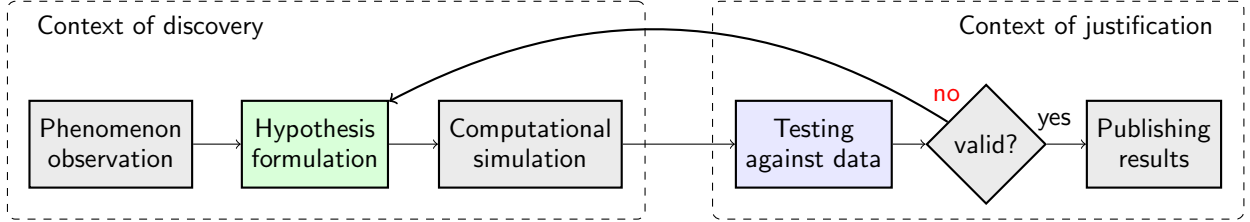


FIG. 1: A view of the scientific method life cycle. It highlights hypothesis formulation and a backward transition to reformulation if predictions ‘disagree’ with observations.

Simulation laboratories provide scientists and engineers with very large, possibly huge datasets that reconstruct phenomena of interest in high resolution. Some examples are the Johns Hopkins Turbulence Databases,¹⁴ and the Human Brain Project (HBP) neuroscience simulation datasets.¹³ A core motivation for the open delivery of such data is enabling new insights and discoveries through *hypothesis testing against observations*.

Nonetheless, while the use case for *exploratory analytics* is well understood and many of its challenges have already been coped with so that high-resolution simulation data are increasingly more accessible,^{1,2} only very recently the use case of hypothesis management has been taken into account for *predictive analytics*.⁹ There is a pressing call for innovative technology to integrate (observed) data and (simulated) theories in a unified framework.⁶

Once parametrized access to large-scale simulation data is delivered, tools for connecting hypothesis formulation and testing into the data-driven science pipeline could open promising possibilities for the scientific method at industrial scale. In fact, the point has just been raised by leading neuroscientists in the context of the HBP, who are incisive on the compelling argument that massive simulation databases should be constrained by experimental data in corrective loops to test precise hypotheses.⁸ (p. 28)

Fig. 1 shows a simplified view of the scientific method life cycle. It distinguishes the phases of exploratory analytics (context of discovery) and predictive analytics (context of justification), and highlights the loop between the hypothesis formulation and testing stages. In this article we address the gap currently separating these two stages of the scientific method in the context of data-driven science. We present a synthesis method and tool, named Υ -DB, for enabling data-driven hypothesis management and predictive analytics in a probabilistic database. It has been demonstrated for simulation data generated from ODE-physiological models,¹⁰ which are available at the Physiome open simulation laboratory.³

TABLE I: Simulation data management vs. hypothesis data management.

| Simulation data management | Hypothesis data management |
|-----------------------------------|------------------------------------|
| Exploratory analytics | Predictive analytics |
| Raw data | Sample data |
| Extremely large (TB, PB) | Very large (MB, GB) |
| Dimension-centered access pattern | Claim-centered access pattern |
| Denormalized for faster retrieval | Normalized for uncertainty factors |
| Incremental-only data updates | Probability distribution updates |

I. HYPOTHESIS DATA MANAGEMENT

Challenges for enabling an efficient access to high-resolution, raw simulation data have been documented from both supercomputing,¹⁴ and database research viewpoints;¹⁶ and pointed as key to the use case of exploratory analytics. The extreme scale of the raw data has motivated such non-conventional approaches for data exploration, viz., the ‘immersive’ query processing (move the program to the data),¹⁴ or ‘in situ’ query processing in the raw files.¹⁶ Both exploit the spatial structure of the data in their indexing schemes.

Simulation data, nonetheless, being generated and tuned from a combination of theoretical and empirical principles, has a distinctive feature to be considered when compared to data generated by high-throughput technology in large-scale scientific experiments such as in astronomy and particle physics surveys. It has a pronounced *uncertainty* component that motivates the use case of hypothesis data management for *predictive analytics*.⁹ Essential aspects of hypothesis data management can be described in contrast to simulation data management as follows — Table I summarizes our comparison.

- *Sample data.* Hypothesis management shall not deal with the same volume of data as in simulation data management for exploratory analytics, but with samples of it. This is aligned, for example, with the architectural design of CERN’s particle-physics experiment and simulation ATLAS, where there are four tier/layers of data. The volume of data significantly decreases from (tier-0) the raw data to (tier-3) the data actually used for analyses such as hypothesis testing.² (p. 71-2) Samples of raw simulation data are to be selected for comparative studies involving competing hypotheses in the presence of evidence (sample observational data). This principle is also aligned with how data are delivered at model repositories. Since observations are usually less available,

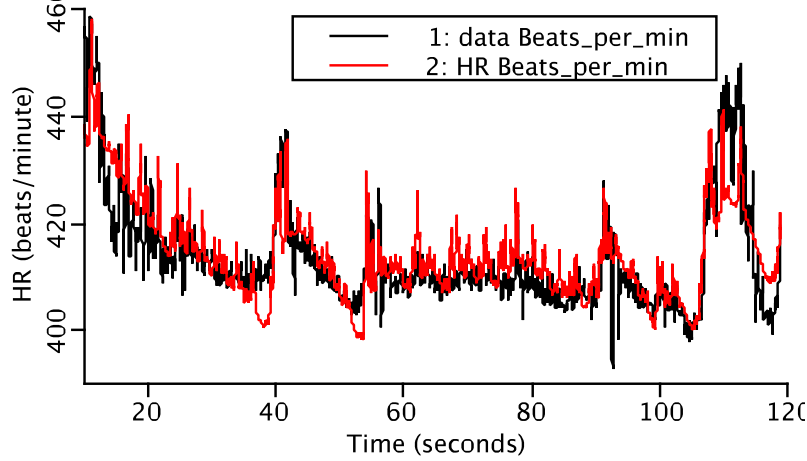


FIG. 2: Theoretical data generated from a baroreflex model (‘HR’ in red color) for Dahl SS Rat and its target observations (‘data’ in black). (source: Bugenhagen et al.⁵).

only the fragment (sample) of the simulation data that matches in coordinates the (sample) of observations is required out of simulation results for comparative analysis. For instance, the graph shown in Fig. 2 from the Virtual Physiological Rat Project (VPR1001-M) aligns simulation data (heart rates from a baroreflex model) with observations on a Dahl SS rat strain.⁵ Here, the simulation is originally set to produce predictions in the time resolution of 0.01. But since the observational sample is only as fine as 0.1, the predicted sample is rendered to match the latter in this particular analysis.

- *Claim-centered access pattern.* In simulation data management the access pattern is dimension-centered (e.g., based on selected space-time coordinates) and the data are denormalized for faster retrieval, as typical of Data Warehouses (DW’s) and On-Line Analytical Processing (OLAP) applications for decision making — in contrast to On-Line Transaction Processing (OLTP) applications for daily operations and updates. In particular, on account of the so-called ‘big table’ approach, each state of the modeled physical system is recorded in a large, single row of data. This is fairly reasonable for an Extract-Transform-Load (ETL) data ingestion pipeline characterized by batch-, incremental-only updates. Such a setting is in fact fit for exploratory analytics, as entire states of the simulated system shall be accessed at once (e.g., providing data to a visualization system). Altogether, data retrieval is critical and there is no risk of update anomalies. Hypothesis management, in contrast, should be centered on

claims identified within the hypothesis structure by means of data dependencies. Since the focus is on resolving uncertainty for decision making (which hypothesis is a best fit?), the data must be normalized based on *uncertainty factors*. This is key for the correctness of uncertainty modeling and efficiency of probabilistic reasoning, say, in a probabilistic database.¹⁷ (p.30-1)

- *Uncertainty modeling.* In uncertain and probabilistic data management,¹⁷ the uncertainty may come from two sources: *incompleteness* (missing data), and *multiplicity* (inconsistent data). Hypothesis management on sample simulation data is concerned with the multiplicity of prediction records due to competing hypotheses targeted at the same studied phenomenon. Such a multiplicity naturally gives rise to a probability distribution that may be initially uniform and eventually conditioned on observations. Conditioning is an applied *Bayesian inference* problem that translates into database update for transforming the prior probability distribution into a posterior.⁹

Overall, hypothesis data management is also OLAP-like, yet markedly different from simulation data management.

A key point that distinguishes hypothesis management is that a fact or unit of data is defined by its **predictive content**, not only by its dimension coordinates. Every clear-cut prediction is a claim identified on account of available dependencies. Accordingly, the data should be decomposed and organized for a claim-centered access pattern.

In what follows we present the use case of hypothesis management and predictive analytics by an example extracted from the Physiome open simulation laboratory (<http://www.physiome.org>).

II. USE CASE: COMPUTATIONAL PHYSIOLOGY HYPOTHESES

Judy is a researcher in computational physiology who got a set of observations of hemoglobin oxygen saturation from the literature in order to test three different theoretical models stored in an open simulation lab (OSL) against it. She knows about Υ-DB, a tool recently plugged into the OSL for data-driven hypothesis management and testing,

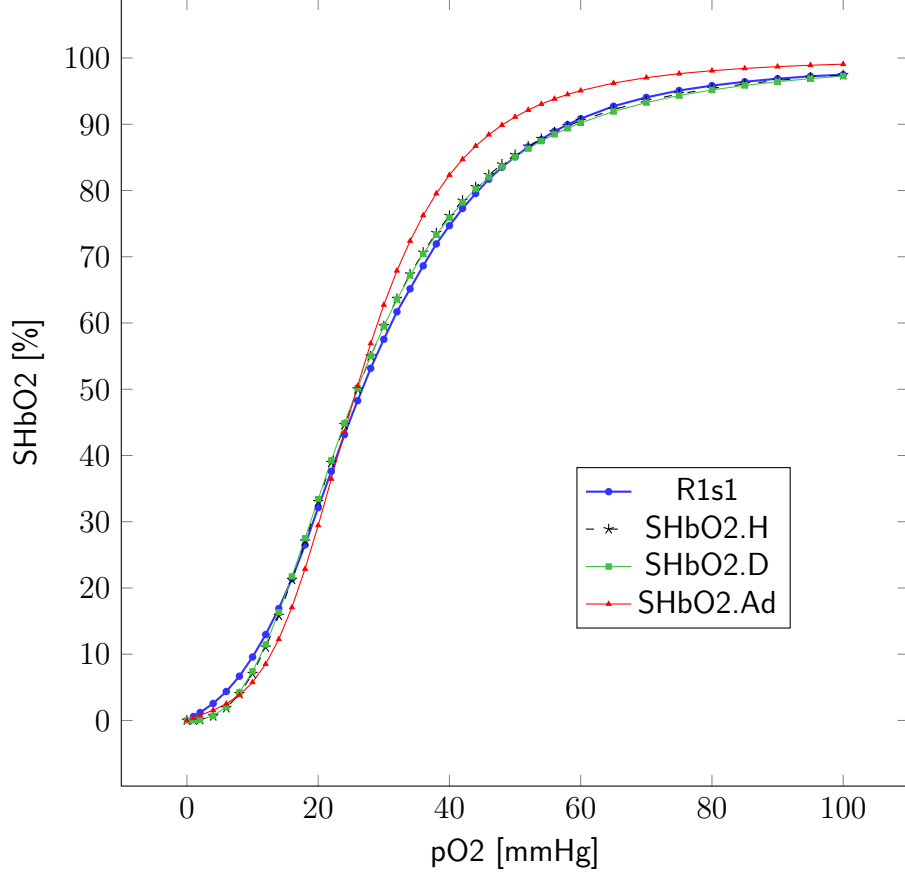


FIG. 3: Hemoglobin oxygen saturation hypotheses (SHbO2.{H, D, Ad} curves) and their target observations (‘R1s1’ dataset). (source: Physiome).

and decides to try it in view of refining her assessment and reporting, otherwise based on visual analytics only (see Fig. 3). As the models are particularly small, she is able to note by herself that the simplest model is Hill’s Equations for O_2 binding to hemoglobin (Eqs. 1) and then wonders whether it could turn out to be the fittest hypothesis. Fig. 4 shows the textual meta-data easily provided by Judy into the Υ -DB system about her study — special attribute symbols ϕ and v are (resp.) identifiers for the phenomenon and the hypotheses.

$$\begin{cases} \text{SHbO2} &= \text{KO2} \cdot \text{pO2}^n / (1 + \text{KO2} \cdot \text{pO2}^n) \\ \text{KO2} &= \text{p50}^{(-n)} \end{cases} \quad (1)$$

| HYPOTHESIS | v | Name | Description |
|------------|-----|-----------|---|
| | 28 | HbO.Hill | Hill’s Equation for O2 binding to hemoglobin. |
| | 31 | HbO.Adair | Hemoglobin O2 saturation curve using Adair’s 4-site equation. |
| | 32 | HbO.Dash | Hemoglobin O2 saturation curve at varied levels of PCO2 and pH. |

| PHENOMENON | ϕ | Description |
|------------|--------|--|
| | 1 | Hemoglobin oxygen saturation with observational dataset from Sevenringhaus 1979. |

FIG. 4: Description of Judy’s hypotheses with their original id’s from Physiome’s OSL.

III. HYPOTHESIS ENCODING

Scientific hypotheses are tested by way of their predictions. In the form of mathematical equations like Eqs. 1, hypotheses symmetrically relate aspects of the studied phenomenon. For computing predictions, however, hypotheses are used asymmetrically like *functions*.¹⁵ They take a given valuation over input variables (parameters) to produce values of output variables (the predictions). Interestingly, such an asymmetry can be detected automatically to establish functional dependencies that unravel the structure of the predictive data.¹¹

We abstract a system of mathematical equations into a structural equation model $\mathcal{S}(\mathcal{E}, \mathcal{V})$,¹⁵ where \mathcal{E} is the set of equations and \mathcal{V} the set of all variables appearing in them. For instance, note that Hill’s Eqs. (1) are $|\mathcal{E}| = 2$, with $\mathcal{V} = \{\text{SHbO2}, \text{KO2}, \text{pO2}, \text{n}, \text{p50}\}$. Intuitively they do not form a valid (computational model) structure, as $|\mathcal{E}| \neq |\mathcal{V}|$. They must be completed by setting domain and parameter values. In this specific case, domain function $f_3(\text{pO2})$ and constant functions $f_4(\text{n})$, $f_5(\text{p50})$ are included into Hill’s structure \mathcal{S} .

In view of uncertainty modeling, we need to derive a set Σ_{28} of functional dependencies (as ‘causal’ orientations)¹⁵ from Hill’s structure. We focus on its implicit data dependencies and get rid of constants and possibly complex mathematical constructs. By exploiting Hill’s global structure \mathcal{S} , equation $\text{KO2} = \text{p50}^{(-\text{n})}$, e.g., is oriented towards KO2, which is then a (prediction) variable *functionally dependent* on (parameters) p50 and n. Yet a dependency like $\{\text{p50}, \text{n}\} \rightarrow \{\text{KO2}\}$ may hold for infinitely many equations (think of, say, how many polynomials satisfy that dependency ‘signature’). In fact, we need a way to identify the equation’s mathematical formulation precisely, i.e., an abstraction of its data-

level semantics. This is achieved by introducing hypothesis id v as a special attribute in the functional dependency (see Σ_{28} below; compare it with Eqs. 1).^{9,11} We adopt here the usual notation for functional dependencies from the database literature, without braces.

$$\begin{aligned}\Sigma_{28} = \{ \quad & \text{KO2 } n \text{ pO2 } v \rightarrow \text{SHbO2}, \\ & n \text{ p50 } v \rightarrow \text{KO2}, \\ & \phi \rightarrow n, \\ & \phi \rightarrow \text{p50} \quad \}. \end{aligned}$$

All dependencies containing v in their left-hand sides mean that the right-hand side variable has its values predicted on account of the values of the variables on the left. The other special attribute, the phenomenon id ϕ , appears in dependencies of the form $\phi \rightarrow x$. These are informative that x has been identified a parameter whose value is set empirically. That is, it is set ‘outside’ of the hypothesis model, contributing to the parameter valuation that defines a particular trial on the hypothesis and then grounds it for a target phenomenon.

That is a data representation of the structure of a scientific hypothesis.⁹ The causal reasoning on the global structure \mathcal{S} is a challenging, yet accessible problem. It can be performed by an efficient algorithm.¹¹ As a result, a total ‘causal’ mapping (a bijection) is rendered from set \mathcal{E} of equations to set \mathcal{V} of variables, i.e., every equation is oriented towards exactly one of its variables. This technique is provably very efficient, viz., $O(\sqrt{|\mathcal{E}|} |\mathcal{S}|)$, where $|\mathcal{S}|$ is the total number of variable appearances in all equations, i.e., a measure of how dense the hypothesis is. So far we have tested it in scale for processing hypotheses whose structures are sized up to $|\mathcal{S}| \lesssim 1M$, with $|\mathcal{E}| = |\mathcal{V}| \leq 2.5 K$.¹¹

Our technique for hypothesis encoding relies on the availability of the hypothesis structure (its equations) in a machine-readable format such as W3C’s MathML. In fact, it has been a successful design decision of Physiome’s OSL to standardize the MathML-compliant Mathematical Modeling Language (MML) for model specification and sharing towards reproducibility (cf. <http://www.physiome.org/jsim/>). The Υ -DB system is then provided with an XML wrapper component to extract Physiome’s models encoded in MML and infer its causal dependencies. The same can be done for every OSL if its computational models are structured in declarative form such as in a MathML file.

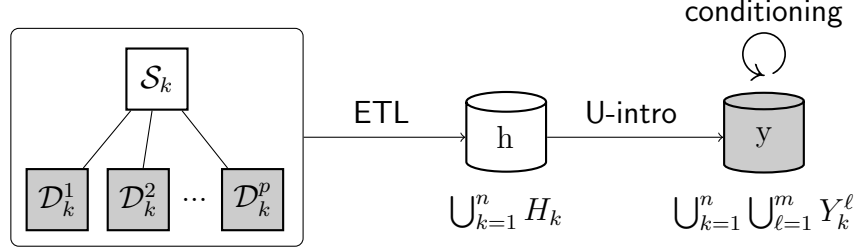


FIG. 5: Synthesis pipeline for processing hypotheses as uncertain and probabilistic data. For each hypothesis k , its structure \mathcal{S}_k is given in a machine-readable format, and all its sample simulation data trials $\bigcup_{i=1}^p \mathcal{D}_k^i$ are indicated their target phenomenon id ϕ and loaded into a ‘big table’ H_k . Then the synthesis comes into play to read a base of possibly very many hypotheses $\bigcup_{k=1}^n H_k$ and transform them into a probabilistic database where each hypothesis is decomposed into claim tables $\bigcup_{\ell=1}^m Y_k^\ell$. A probability distribution is computed for each phenomenon ϕ , covering all the hypotheses and their trials targeted at ϕ . This distribution is then updated into a posterior in the presence of observational data.

IV. SYNTHESIS PIPELINE

When Judy inserts a hypothesis k , into Υ -DB, the system extracts the set of (causal) functional dependencies from its mathematical structure and then she can upload its sample simulation data into a ‘big table’ H_k containing all its variables as relational attributes in the table. For the upload, she chooses a phenomenon to be targeted by the hypothesis simulation data. Both the hypothesis structure and its data are input to the synthesis pipeline shown in Fig. 5.

Normalization of the ‘big table’ (see Fig. 6), as discussed above, is not desirable because its data are not be updated (only re-inserted if necessary). For hypothesis management, however, the uncertainty has to be decomposed/normalized so that the uncertainty of one claim may not be undesirably mixed with the uncertainty of another claim. In fact, we perform further reasoning, viz., acyclic pseudo-transitive reasoning over functional dependencies,¹¹ to process Σ_{28} into another set Σ'_{28} of dependencies. This is ensured to have, for each predictive variable, exactly one dependency with it on the right-hand side and all its uncertainty factors (so-called ‘first causes’) on the left-hand side.¹¹

$$\Sigma'_{28} = \left\{ \begin{array}{l} \text{n p50 } \phi \text{ v } \rightarrow \text{KO2}, \\ \text{n p50 pO2 } \phi \text{ v } \rightarrow \text{SHbO2} \end{array} \right\}$$

| H_{28} | ϕ | v | pO2 | KO2 | SHbO2 | n | p50 |
|----------|--------|-----|-----|---------------------|---------------------|-----|-----|
| | 1 | 28 | 0 | 1.51207022127057E-4 | 0 | 2.7 | 26 |
| | 1 | 28 | 0.1 | 1.51207022127057E-4 | 3.01697581987324E-7 | 2.7 | 26 |
| | 1 | 28 | 0.2 | 1.51207022127057E-4 | 1.96043341970514E-6 | 2.7 | 26 |
| | 1 | 28 | ... | 1.51207022127057E-4 | ... | 2.7 | 26 |
| | 1 | 28 | 100 | 1.51207022127057E-4 | 9.74346796798538E-1 | 2.7 | 26 |

FIG. 6: ‘Big table’ H_{28} of hypothesis $v = 28$ (Hill’s equation) loaded with one sample trial dataset targeted at phenomenon $\phi = 1$. The main predictive variable is hemoglobin oxygen saturation SHbO2, whose values evolve with the values of dimension pO2.

V. PROBABILISTIC DATABASE

It is a basic design principle for uncertainty modeling to define exactly one random variable for each actual uncertainty factor (*u-factor*, for short). The hypothesis model is itself a theoretical u-factor, whose uncertainty comes from the multiplicity of models targeted at explaining the same phenomenon. Additionally, the multiplicity of trials on each hypothesis (alternative parameter settings) targeted at the same phenomenon gives rise to empirical u-factors. In fact, a hypothesis model can only approximate a phenomenon very well if it is properly tuned (calibrated) for the latter. The probability distribution on a phenomenon must take into account both kinds of u-factors to support hypothesis testing in this broad sense. Each u-factor is captured into a random variable in the probabilistic database.

We carry out the algorithmic transformation from each hypothesis ‘big table’ to the probabilistic database through the probabilistic world-set algebra of the U-relational model, an extension of relational algebra for managing uncertain and probabilistic data.^{12,17}

U-relations have in their schema a set of pairs (V_i, D_i) of *condition columns* to map each discrete random variable x_i to one of its possible values (e.g., $x_0 \mapsto 1$). The ‘world table’ W , inspired in pc-tables,¹⁷ stores their marginal probabilities. Fig. 7 shows the probabilistic U-relational tables synthesized for hypothesis $v = 28$. Any row of, say, table Y_{28}^4 , has the same joint probability distribution $\Pr(\theta) \approx .33$, which is associated with possible world $\theta = \{x_0 \mapsto 1, x_1 \mapsto 1, x_2 \mapsto 1\}$. The probabilistic inference is performed in aggregate queries by the **conf** operator based on the marginal probabilities stored in world table W .¹²

Such a probabilistic database should bear desirable design-theoretic properties for uncer-

| | | | | | | | | | | | | | |
|-------|-----------------|--------|-----|--|------------|-----------------|--------|-----|--|------------|-----------------|--------|-------|
| Y_0 | $V \mapsto D$ | ϕ | v | | Y_{28}^1 | $V \mapsto D$ | ϕ | n | | Y_{28}^2 | $V \mapsto D$ | ϕ | $p50$ |
| | $x_0 \mapsto 1$ | 1 | 28 | | | $x_1 \mapsto 1$ | 1 | 2.7 | | | $x_2 \mapsto 1$ | 1 | 26 |
| | $x_0 \mapsto 2$ | 1 | 31 | | | | | | | | | | |
| | $x_0 \mapsto 3$ | 1 | 32 | | | | | | | | | | |

| | | | | | | | |
|------------|-------------------|-------------------|-------------------|--------|-----|---------------------|--|
| Y_{28}^3 | $V_0 \mapsto D_0$ | $V_1 \mapsto D_1$ | $V_2 \mapsto D_2$ | ϕ | v | KO2 | |
| | $x_0 \mapsto 1$ | $x_1 \mapsto 1$ | $x_2 \mapsto 1$ | 1 | 28 | 1.51207022127057E-4 | |

| | | | | | | | | | | | |
|------------|-------------------|-------------------|-------------------|--------|-----|-----|---------------------|--|-----|-----------------|-----|
| Y_{28}^4 | $V_0 \mapsto D_0$ | $V_1 \mapsto D_1$ | $V_2 \mapsto D_2$ | ϕ | v | pO2 | SHbO2 | | W | $V \mapsto D$ | Pr |
| | $x_0 \mapsto 1$ | $x_1 \mapsto 1$ | $x_2 \mapsto 1$ | 1 | 28 | 0 | 0 | | | $x_0 \mapsto 1$ | .33 |
| | $x_0 \mapsto 1$ | $x_1 \mapsto 1$ | $x_2 \mapsto 1$ | 1 | 28 | 0.1 | 3.01697581987324E-7 | | | $x_0 \mapsto 2$ | .33 |
| | $x_0 \mapsto 1$ | $x_1 \mapsto 1$ | $x_2 \mapsto 1$ | 1 | 28 | 0.2 | 1.96043341970514E-6 | | | $x_0 \mapsto 3$ | .33 |
| | $x_0 \mapsto 1$ | $x_1 \mapsto 1$ | $x_2 \mapsto 1$ | 1 | 28 | ... | ... | | | $x_1 \mapsto 1$ | 1 |
| | $x_0 \mapsto 1$ | $x_1 \mapsto 1$ | $x_2 \mapsto 1$ | 1 | 28 | 100 | 9.74346796798538E-1 | | | $x_2 \mapsto 1$ | 1 |

FIG. 7: U-relations synthesized for hypothesis $v = 28$. The model competition on phenomenon $\phi = 1$ is captured into the probability distribution associated with random variable x_0 , see U-relation Y_0 . The ‘world table’ W stores the marginal probabilities on the random variables. Observe that there is no parameter uncertainty (multiplicity) in this hypothesis $v = 28$, where random variables x_1 and x_2 are associated with its parameters n and $p50$. Values of predictive variables are then annotated with the uncertainty coming from their u-factors, which are then combined into a joint probability distribution.

tainty modeling and probabilistic reasoning.¹⁷ (p. 30-1) In fact, it is in Boyce-Codd normal form w.r.t. the (causal) functional dependencies and its uncertainty decomposition (into marginal probabilities) is recoverable by a lossless join (joint probability distribution).¹¹

VI. PREDICTIVE ANALYTICS

Noticeably, the prior probability distribution on ‘explanation’ random variable x_0 is uniform (see world table W in Fig. 7). Now that Judy’s hemoglobin oxygen saturation hypotheses are encoded with their sample simulation data properly stored in the probabilistic database, she is keen to see the results, the hypothesis rating/ranking based on the observed data. The insertion of the latter into Υ -DB is straightforward. It is loaded into a relational table (not shown) from a CSV file and associated with phenomenon $\phi = 1$.

The system then enables her to carry out Bayesian inference steps that update at each step the prior distribution of her interest to a posterior. In such computational science use cases, as we have *discrete* random variables mapped to the possible values of (numerical) prediction variables whose domain are *continuous* (double precision), the Bayesian inference is applied for normal mean (likelihood function) with a discrete prior (probability distribution).⁴

| STUDY | ϕ | v | pO2 | SHbO2 | Prior | Posterior |
|-------|--------|-----|-----|----------------------|-------|-----------|
| | 1 | 32 | 1 | 0.178973375779681E-3 | .333 | .335441 |
| | 1 | 28 | 1 | 0.151184162020125E-3 | .333 | .335398 |
| | 1 | 31 | 1 | 3.789100566457180E-3 | .333 | .329161 |
| | ... | ... | ... | ... | ... | ... |
| | 1 | 32 | 100 | 9.72764121981342E-1 | .333 | .335441 |
| | 1 | 28 | 100 | 9.74346796798538E-1 | .333 | .335398 |
| | 1 | 31 | 100 | 9.90781330988763E-1 | .333 | .329161 |

FIG. 8: Results of Judy’s analytics on the computational physiology hypotheses. The predictions from each hypothesis are aligned with the observational dataset which is of smaller resolution. The hypothesis rating/ranking is derived from Bayesian inference.

The procedure uses normal density function (Eq. 2) with standard deviation σ to compute the likelihood $f(y | \mu_k)$ for each competing prediction μ_k given observation y . But as we actually have a sample of independent observed values y_1, \dots, y_n (viz., measured hemoglobin oxygen saturations SHbO2 over different oxygen partial pressures pO2). Then, the likelihood $f(y_1, \dots, y_n | \mu_k)$ for each competing trial μ_k , is computed as a product $\prod_{j=1}^n f(y_j | \mu_{kj})$ of the single likelihoods $f(y_j | \mu_{kj})$.⁴ Bayes’ rule is then settled by (Eq. 3) to compute the posterior $p(\mu_k | y_1, \dots, y_n)$ given prior $p(\mu_k)$.

$$f(y | \mu_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{1}{2\sigma^2}(y-\mu_k)^2} \quad (2)$$

$$p(\mu_k | y_1, \dots, y_n) = \frac{\prod_{j=1}^n f(y_j | \mu_{kj}) p(\mu_k)}{\sum_{i=1}^m \prod_{j=1}^n f(y_j | \mu_{ij}) p(\mu_i)} \quad (3)$$

Fig. 8 shows the results of Judy’s analytical inquiry into the three hemoglobin oxygen saturation hypotheses given the observations she managed to take from the literature. Unlike her expectations w.r.t. the principle of Occam’s razor, Hill’s model has been beaten by Dash’s model, which is structurally more complex (viz., $|\mathcal{S}_H| = 10$, $|\mathcal{S}_D| = 35$). This top-ranked hypothesis includes additional observables such as pCO₂ and pH. Hypothesis management can provide this and other model statistics (e.g., predictive power, mean computation time, etc), which may provide useful metrics to assess hypotheses qualitatively as well.

The Υ -DB system can also be used to study a single hypothesis under very many al-

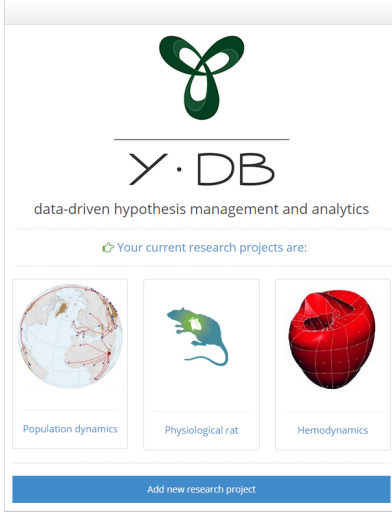
ternative parameter settings (trials) aimed at finding the best fit for a phenomenon. For example, we have applied it for the case of the VPR’s baroreflex hypothesis (Fig. 2),⁵ sized $|\mathcal{S}| = 171$, to find the best fit among $1K$ trials stored in the probabilistic database.¹¹ (p. 81-2)

Altogether, it is worthwhile highlighting that Υ -DB does not provide any new statistical tool for hypothesis testing. It rather can implement a suit of existing tools for enabling data-driven hypothesis management in a systematic fashion on top of state-of-the-art database technology.

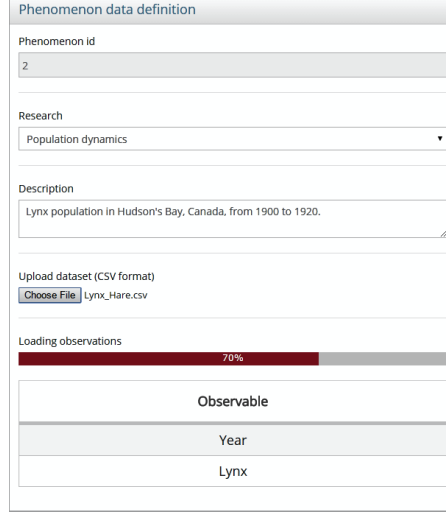
VII. PROTOTYPE SYSTEM

A first prototype of the Υ -DB system has been implemented as a Java web application,¹⁰ with the pipeline component in the server side on top of **MayBMS** (a backend extension of **PostgreSQL**).¹² Fig. 9 shows screenshots of the system in a population dynamics scenario comprising the Malthusian model, the logistic equation and the Lotka-Volterra model applied to predict the Lynx population in Hudson’s Bay in Canada from 1900 to 1920. The observations, collected from Elton and Nicholson,⁷ are used to rank the competing hypotheses and their trials accordingly.

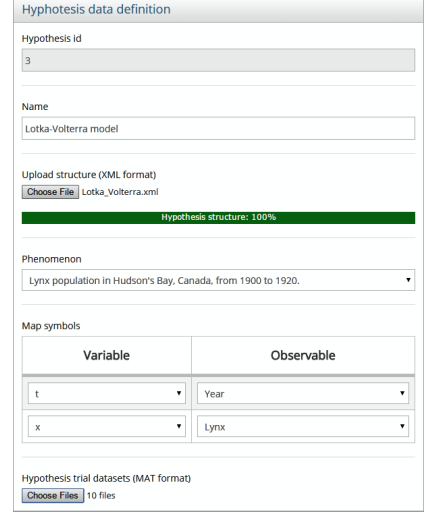
Fig. 9(a) shows the research projects currently available for a user. Figs. 9(b, c) show the ETL interfaces for phenomenon and hypothesis data definition (by synthesis), and then the insertion of hypothesis simulation trial datasets. Note that it requires simple phenomena description, hypothesis naming and file upload to get phenomena and hypotheses available in the system to be managed as probabilistic data. Fig. 9(d) shows the interface for a basic retrieval of simulation data, given a selected phenomenon and a hypothesis trial. Figs. 9(e, f) show two tabs of the predictive analytics module. Note that the user chooses a phenomenon for study and imposes some selectivity criteria onto its observational sample. The system then lists in the next tab the corresponding predictions available, ranked by their probabilities conditioned on the selected observations. In this case, Lotka-Volterra’s model (under trial $\text{tid} = 2$) is the top-ranked hypothesis to explain the Lynx population observations in Hudson’s Bay from 1900 to 1920.



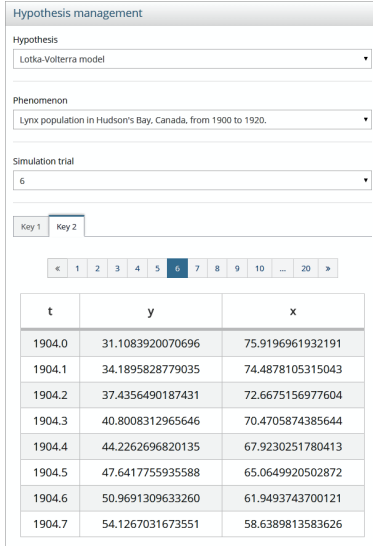
(a) Research dashboard.



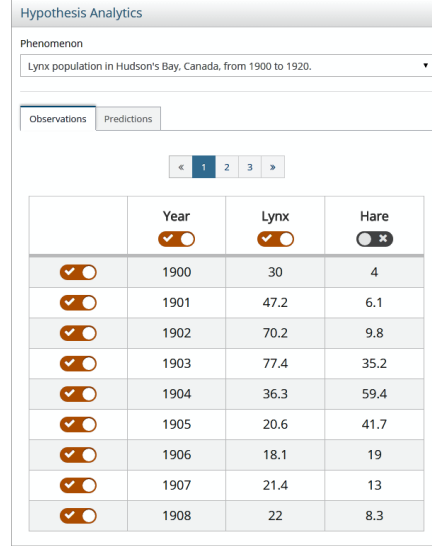
(b) Phenomenon data def.



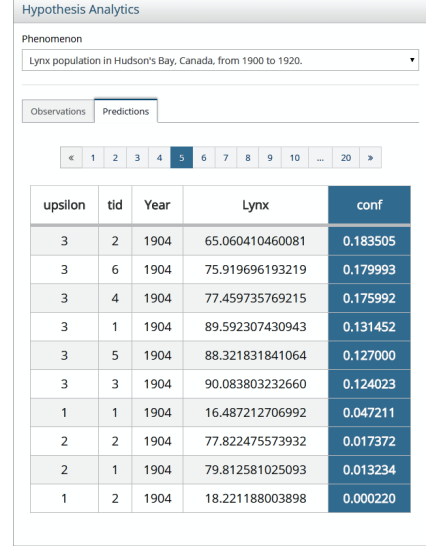
(c) Hypothesis data definition.



(d) Hypothesis data retrieval.



(e) Selected observations tab.



(f) Ranked predictions tab.

FIG. 9: Screenshots of the Y-DB system prototype.

VIII. CONCLUSIONS

Hypothesis data management is a promising new research field towards taking more value out of the theoretical data available in open simulation laboratories. Our work is a first effort to define its use case in the context of simulation data management. It proposes core principles and techniques for encoding and managing hypotheses as uncertain and probabilistic data,^{9,11} enabling data-driven hypothesis testing and predictive analytics.¹⁰

A next step is to apply the method to hypotheses that are complex not only in terms of

number of equations and coupled variables, but also dimensionally like in PDE models of fluid dynamics. Besides, major directions of future work are (i) to improve the statistical capabilities of the Υ -DB system for supporting the data sampling out of simulation results, and (ii) to push its scalability forward to allow for hypothesis testing on samples of larger scale.

IX. ACKNOWLEDGMENTS

This work has been supported by the Brazilian funding agencies CNPq (grants n^o 141838/2011-6, 309494/2012-5) and FAPERJ (grants INCT-MACC E-26/170.030/2008, ‘Nota 10’ E-26/100.286/2013). We thank IBM for a Ph.D. Fellowship 2013-2014.

REFERENCES

- ¹Y. Ahmad, R. Burns, M. Kazhdan, C. Meneveau, A. Szalay, and A. Terzis. Scientific data management at the Johns Hopkins Institute for Data Intensive Engineering and Science. *SIGMOD Record*, 39(3):18–23, 2010.
- ²A. Ailamaki, V. Kantere, and D. Dash. Managing scientific data. *Comm. ACM*, 53(6):68–78, 2010.
- ³J. B. Bassingthwaighe. Strategies for the Physiome Project. *Ann. Biomed. Eng.*, 28:1043–58, 2000.
- ⁴W. M. Bolstad. *Introduction to Bayesian Statistics*. Wiley-Interscience, 2nd edition, 2007.
- ⁵S. M. Bugenhagen, A. W. J. Cowley, and D. A. Beard. Identifying physiological origins of baroreflex dysfunction in salt-sensitive hypertension in the Dahl SS rat. *Physiological Genomics*, 42:23–41, 2010.
- ⁶J. B. Cushing. Beyond big data? *Computing in Science & Engineering*, 15(5):4–5, 2013.
- ⁷C. Elton and M. Nicholson. The ten-year cycle in numbers of the lynx in Canada. *Journal of Animal Ecology*, 11(2):215–44, 1942.
- ⁸Y. Frégnac and G. Laurent. Where is the brain in the Human Brain Project? *Nature*, 513:27–9, 2014.
- ⁹B. Goncalves and F. Porto. Υ -DB: Managing scientific hypotheses as uncertain data. *PVLDB*, 7(11):959–62, 2014.

- ¹⁰B. Goncalves, F. C. Silva, and F. Porto. Y-DB: A system for data-driven hypothesis management and analytics. Technical report, LNCC, 2015. (available at CoRR abs/1411.7419).
- ¹¹B. Gonçalves. *Managing large-scale scientific hypotheses as uncertain and probabilistic data*. PhD thesis, National Laboratory for Scientific Computing (LNCC), Brazil, 2015. (available at CoRR abs/1501.05290).
- ¹²C. Koch. *MayBMS: A system for managing large uncertain and probabilistic databases*. In C. Aggarwal (ed.), *Managing and Mining Uncertain Data*, Chapter 6. Springer-Verlag, 2009.
- ¹³H. Markram. The Blue Brain Project. *Nature Reviews Neuroscience*, 7:153–60, 2006.
- ¹⁴E. Perlman, R. Burns, Y. Li, and C. Meneveau. Data exploration of turbulence simulations using a database cluster. In *Proc. of ACM/IEEE Supercomputing (SC’07)*, 2007.
- ¹⁵H. Simon and N. Rescher. Cause and counterfactual. *Philosophy of Science*, 33(4):323–40, 1966.
- ¹⁶A. Stougiannis, F. Tauheed, M. Pavlovic, T. Heinis, and A. Ailamaki. Data-driven Neuroscience: Enabling breakthroughs via innovative data management. In *Proc. of the International Conference on Management of Data (SIGMOD ’13)*, 2013.
- ¹⁷D. Suciu, D. Olteanu, C. Ré, and C. Koch. *Probabilistic Databases*. Morgan & Claypool Publishers, 2011.

Bernardo Gonçalves is currently a postdoctoral researcher at the University of Michigan, Ann Arbor. His research centers on the design of data systems and analytics to support the scientific method at scale. He holds a Ph.D. in Computational Modeling from the National Laboratory for Scientific Computing (LNCC) in Brazil, and a M.Sc. and a B.Sc. in Computer Science from the Federal University of Espirito Santo (UFES) in Brazil.

Fabio Porto is a researcher at LNCC, in Brazil. He holds a PhD in Computer Science from PUC-Rio, in a sandwich program at INRIA, France. Prior to LNCC, he worked as a researcher at EPFL, Database Laboratory, in Switzerland. At LNCC, he coordinates the Data Extreme Lab, whose focus is on research and development activities involving the design of techniques, algorithms and models for scientific data management and analysis. He is a member of the Brazilian Computer Society (SBC) and Association of Computing Machinery (ACM).